Transformer Network Intuition

- Vanishing gradients

RNN → GRU → LSTM

→

sequencial models, increased Complexity

Has to compute token by token. each unit / token in these networks is a bottleneck to the flow of information.

Base Idea: Attention + CNN

RNN: $y^{<0>}$ $y^{<1>}$ $y^{<2>}$ ...

CNN:

parallel

Self-Attention

Self-Attention

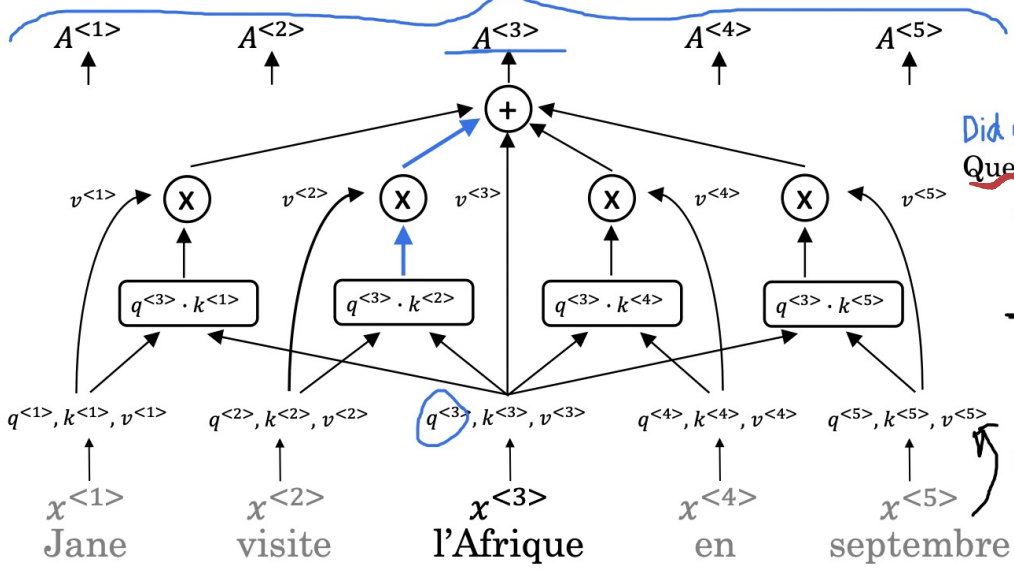
vectorized for Q.

$$A(q,K,V) = \sum_i \frac{\exp(e^{<q \cdot k^{<i>}>})}{\sum_j \exp(e^{<q \cdot k^{<j>}>})} v^{<i>}$$

Softmax

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

scaled dot product, preventing exploding



asked Question

Did what?

Query (Q) Key (K) Value (V)

$q^{<1>}$ $k^{<1>}$ $v^{<1>}$

$q^{<2>}$ $k^{<2>}$ $v^{<2>}$

$q^{<3>}$ $k^{<3>}$ $v^{<3>}$

$q^{<4>}$ $k^{<4>}$ $v^{<4>}$

$q^{<5>}$ $k^{<5>}$ $v^{<5>}$

person action

Jane visit

What's happening there?

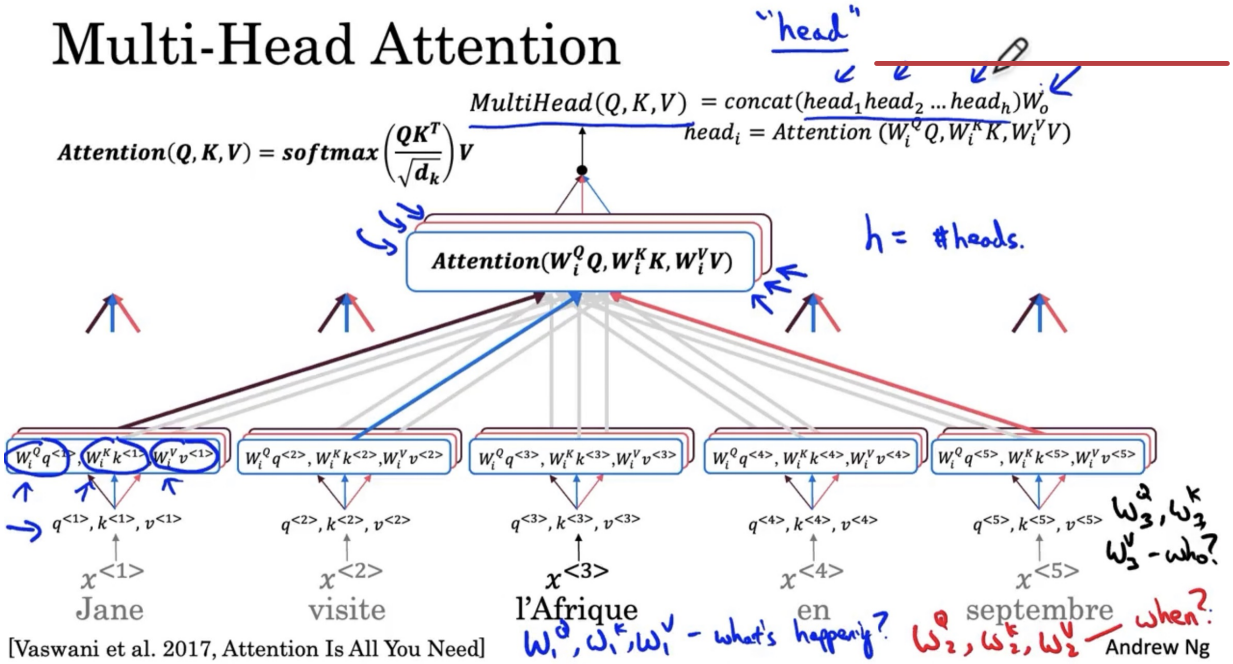
W^Q, W^K, W^V

[Vaswani et al. 2017, Attention Is All You Need]

scaled dot product attention.

Andrew Ng

Multi-Head Attention



Transformer Network